Your Diffusion Model is Secretly a Zero-Shot Classifier

Abstract

The recent wave of large-scale text-to-image diffusion models has dramatically increased our text-based image generation abilities. These models can generate realistic images for a staggering variety of prompts and exhibit impressive compositional generalization abilities. Almost all use cases thus far have solely focused on sampling; however, diffusion models can also provide conditional density estimates, which are useful for tasks beyond image generation. In this paper, we show that the density estimates from large-scale text-to-image diffusion models like Stable Diffusion can be leveraged to perform zero-shot classification without any additional training. Our generative approach to classification, which we call Diffusion Classifier, attains strong results on a variety of benchmarks and outperforms alternative methods of extracting knowledge from diffusion models. Although a gap remains between generative and discriminative approaches on zero-shot recognition tasks, our diffusion-based approach has stronger multimodal compositional reasoning abilities than competing discriminative approaches. Finally, we use Diffusion Classifier to extract standard classifiers from class-conditional diffusion models trained on ImageNet. These models approach the performance of SOTA discriminative classifiers and exhibit strong "effective robustness" to distribution shift. Overall, our results are a step toward using generative over discriminative models for downstream tasks. Results and visualizations on our website: diffusion-classifier.github.io/